COMPUTERIZED DATA SHARING: THE SARG EXAMPLE OF COOPERATIVE RESEARCH

Sylvia W. Gaines
Rachel Most

Department of Anthropology
Arizona State University
Tempe, Arizona 85287
U.S.A.

The concept of computerized data sharing is relatively new to American archaeology. An example which will serve to illustrate this approach is a cooperative project known as the Southwestern Anthropological Research Group (SARG). However, before we turn to this example, let us outline some of the general benefits and problems of data sharing we have found from our experience.

In this time of rising costs and shrinking research budgets expenses are becoming an increasingly important factor in archaeological investigations. It is far less expensive for a number of archaeologists to share a common data base on a single computer facility. Equally important is the broader base of research made possible by such an approach. Pooling a number of data sets will result in a more extensive data base than any one individual or field project could hope to acquire. Such an expanded information source provides a more in depth basis for research questions requiring either survey or excavation data. Underlying the idea of shared data is the recognition by the participating archaeologists of some minimal level of common research objectives. Our hope is that, as the utility of such an approach is demonstrated, more researchers and potential contributors will design their own projects to allow them, at least in part, to incorporate information into the larger, data sharing programs. Cooperative projects of this nature require a more rigorous consideration of terms, definitions and data standards. This latter factor alone will have far reaching benefits for archaeology.

American archaeologists are deeply concerned with the ever increasing human impact on the cultural resources of their country. Guidelines regarding data recovery and strict environmental laws do little more than mitigate the problem of this rapidly disappearing resource. As our archaeological sites diminish, we must turn more and more to existing data sources for research and analysis, and the role of automated data bases becomes increasingly important. Computerized data sharing projects may indeed establish data bases which will provide the information for future archaeological research.

Computerized data sharing and cooperative projects are not without their problems. Since computerized data sharing is still in a pioneering stage there has been little precedence for organizational and procedural guidelines. The degree of cooperation can vary considerably among projects but the volunteer nature and loose organization of most data sharing projects place constraints on the research design. Basic issues of who submits data and who manages the data base are critical to the organization and operation of a data sharing venture and must be faced forthright and the decision supported.

A concern highlighted by our data sharing projects is the issue of data comparability and standardization. Advocates of comparability hold that it must be achieved at the level of inference and test results and not at the observational level. Others argue that standardization is the starting point and comparability cannot be achieved without this form of quality control.

Cultural resource management projects, which clearly dominate American field archaeology today, often involve multiple federal and state agencies. Coventures such as these are particularly susceptable to the constraints imposed by data sharing.

At a 1977 meeting in Arizona, federal and state agencies explored the problem of how to organize and share the increasing amounts of information associated with archaeological investigations. It was noted that each land managing agency was developing its own management information system. These were sufficiently different to make a single data base impractical. A number of different data sets were reviewed to determine similar data categories and only three data elements within all the agencies' systems were identified as common. The categories were: site number, location and county, with the recording institution taken by default from the form heading. Artifact information varied widely; however, whether artifacts had been collected was supplied in all cases. This example is cited to emphasize some of the issues which need to be resolved before any pooling of data can be considered. This example is not meant to infer that such sharing is not possible but that mutual effort early in the definition phase is required by personnel dedicated to a common objective.

A rising concern has been the security of computerized data bases which often contain sensitive locational information. User authorization must be strictly enforced. A data coordinator or data manager is not only desirable but necessary to oversee security measures as they relate to data entry. All data sharing projects need to recognize the desirability of strong safeguards and security measures.

Let us now turn to the Southwestern Anthropological Research Group (SARG) for a more specific example to illustrate the data sharing concept. SARG was founded in 1971 as a cooperative project involving a number of Southwestern archaeologists interested in devoting a portion of their individual research to problems of broad cultural significance. This volunteer organization is composed of 15-20 active researchers who are located in institutions

and agencies throughout the country, but whose main fieldwork centers in the American Southwest. For the past ten years we have been collecting survey data in a standardized format, storing these in a common computerized data base at Arizona State University, and using the data to address questions relating to site locating behavior (Gaines and Plog 1980:1).

At an early state in this project computer procedures and techniques were developed to handle the extensive data base. From the onset we realized that, in a long term effort of this nature, research focus would undoubtably require modification. Thus our computer methodology was designed with this type of evolution in mind and the data format and processing procedures were established so that these could serve as a basis for a flexible and expanding system for storing and manipulating data (Gaines 1978:121). Let us look at three broad contributions which have resulted from the SARG effort.

Our computerized data recording format which was initiated in 1971, not only served to structure SARG data but has had considerable impact outside this organization. We feel quite certain that the current high standards and more complete observations which generally characterize Southwestern archaeological surveys, are due largely to the SARG survey format (Gaines and Plog 1980:3). In this respect, SARG has served as a model for a number of other institutions considering similar research. The second contribution concerns the growth potential of our data base. There are approximately 20,000 sites which are currently available for data entry and, given the level of archaeological work in the Southwest, information on several thousand additional sites per year could be incorporated if we so choose. Obviously, this organization has the potential of amassing a computerized data base of survey information unparalleled in extent and sophistication in American archaeology. Finally, the utili-

zation of a pooled data base on a single computer facility, accessible to data sharing participants who are located throughout the country, offers substantial benefits in terms of cost effectiveness of processing and storage, and the added benefit of having available an extensive data base for research.

Let us now turn to the SARG example and the five topics which we consider key to the organization of a project of this nature - these include the selection of variables, data recording, data entry and verification, data storage and maintenance, and data analysis. Each topic is addressed in terms of the early applications (from 1971 to 1976) as well as the current directions (from 1980 to 1982). (For an expanded discussion of these topics the reader is directed to Gaines and Gaines 1981.)

VARIABLE SELECTION

In the initial years a number of key variables were identified by SARG participants. These variables were based on the research objectives which focused on critical environmental resources as well as social factors (Plog and Hill 1971). As some of the members of SARG requested variables which extended beyond the specific goals of the research design, the artifact, features and structures categories were expanded. By 1975, the data base included information on a maximum of 135 variables from approximately 2400 sites from Arizona, Utah, and New Mexico.

Since 1979 the research focus of SARG has changed with a new emphasis on site locating behaviors as these related to stress situations. This approach resulted in the development of new hypotheses as well as new spatial and temporal limitations. Only data from the post-Archaic, prehistoric Plateau Southwest were to be considered. Certain variables previously included in the SARG

format were deemed unnecessary as well as time consuming to record. These revisions resulted in the reduction of the number of variables from 135 to 74. This process included deleting variables, adding new variables, or slightly modifying the definition of existing variables. For example, site location was previously measured and recorded in longitude and latitude. In the new SARG format, site location is measured according to the Universal Transverse Mercator system (UTM). UTM locations are easily taken from most United States Geological Survey maps. This modification in recording site locations has allowed us to interface our data base with many of the mapping programs available to us.

Another change in variables was a shift from interval to ordinal data. The justification for this decision was that interval variables were too specific for the goals of the SARG research. For example, distance to arable land, measured in meters, clearly taxed the nature of inference that was being attempted. Ordinal scale preserves the information in far more realistic form.

There are both advantages and disadvantages to any classification scheme. A problem common to many classification systems is there are always unique cases where the options offered to the recorder do not fit his or her data. It is critical that the categories and classification systems are open ended. This is especially true in a data sharing situation such as SARG where there is considerable variability among project areas. There are, however, several data recording alternatives. A variable can be coded as "other" which may have consistent meaning for a single participant but would have little analytical value for pooled data. A better alternative is to create a new value for the variable in question. For example, a common plant in the Grand Canyon area is agave, yet no code existed for that plant within our vegetation variable. The coding of this plant in any other manner

would have lessened the analytical potential, thus justifying a new code for this plant. The flexibility of an open-ended classificatory system and available computer programs allowed this procedure to be easily accomplished.

Some of the new variables such as "on site landform profile" have allowed more precision in analyses. Previously, we had recorded landform in the traditional Hammond broad categories such as mountain, hill, knoll, etc. Currently, a four-way slope indicator is used to provide a three dimensional portrayal of the landform associated with the site. This new method of recording landforms also allows for greater comparability among the data sets.

In summary, the current variable list is much more efficient than it was seven years ago. The 74 variables provide a great deal of information and are more appropriate for the level of analyses being conducted. It is through exploratory data analyses that the potential of these variables is being recognized.

DATA RECORDING

The method for recording data in the early years of our research required key punched cards or magnetic tape. Each site required five cards to incorporate all necessary data. It was the responsibility of each SARG participant to submit his or her data to Arizona State University. During this time no standard SARG recording form was required. Quality control was the responsibility of each participant.

In 1980 data recording changed radically as a consequence of new data entry procedures. At this time a four page optional form was developed for SARG use. Members have the choice of either coding their data directly onto the SARG form or submitting their field forms. In the latter case, the SARG

staff is responsible for transposing the data onto the SARG forms for data entry. In all cases any data taken from USGS maps are coded by the SARG staff. This information includes variables such as UTM, landform profiles, elevation and vertical relief. To accomplish this map work, a map with sites plotted must be supplied by the SARG member. This procedure has resulted in increased comparability among the data sets. As Gaines and Gaines point out:

> "The key point is that to assure the data base integrity and data quality, standards had to be set for codification, format and variables. While classifications must remain open ended, redundancy should be avoided. This implies a centralized administrative control to be successful" (1981:5).

Two of the key problems with data recording in a shared environment are (1) missing data, and (2) misinterpreting the meaning of the variables. For example, if information such as site date or site location are not recorded, the data set cannot be used in many of the analyses currently being undertaken by the SARG staff. In most instances these data categories have been supplied by participants. With regard to misinterpreted variable definitions, an example which illustrates this point, concerns the variable "room/pithouse count". In the process of conducting analyses on room counts it was realized that some SARG participants record kivas, a ceremonial feature, as rooms while others do not. Thus, some members would code a site with ten surface rooms and one kiva as "11 rooms" while others would code the variable with a "ten".

Although these are common problems which arise in any attempt in cooperative data sharing, SARG has minimized many of them in the past decade of work by

striving for more explicit definitions and a closer communication among the participants.

DATA ENTRY AND VERIFICATION

Originally the SARG data were entered into the UNIVAC 1110 system on keypunched cards. Although each SARG participant was responsible for data accuracy, additional verification was performed once the data were stored on disk. These checks involved performing standard SPSS descriptive statistics on selected variables. Errors were also detected by scanning a printout of the entire data set. However, since there was no accompanying recording form, further checks were not possible.

Installation of a new computer facility (IBM 3081) at Arizona State University required new procedures for data entry. Cards were no longer acceptable, and data were entered from CRT terminals, two of which are located in our archaeology computer laboratory. This transition occurred simultaneously with the change in the SARG format and the initiation of our standardized recording form. Approximately eight months were spent reformating data, acquiring missing data and correcting previous coding errors. Three of the earlier data sets were updated by adding new site information and four new data sets were added. This expanded data base, which now includes approximately 3500 sites, will assure better coverage of the Southwest plateau. Data is still checked manually by viewing a listing on the terminal screen and by "proof-reading" a printout of the entered data to confirm that it corresponds to the forms. However, data errors are most commonly found at a later time during analysis.

Coding and typing errors will no doubt always be present but with the current system they are greatly minimized. Until we implement a system of automatic data verification this problem will remain.

DATA STORAGE AND MAINTENANCE

No funds were available when SARG first began as a cooperative effort. This necessitated the use of a software package system which was available and maintained at the Arizona State University computer center. SPSS (Statistical Package for the Social Sciences) was selected as it had capabilities for both data management and analysis (Gaines 1978:122).

In the earlier years, data were stored on an SPSS SAVEFILE which was later eliminated in favor of independent data sets. Data stored in an SPSS SAVEFILE did not permit interfacing with other program packages nor could the data be viewed on a CRT screen. With the newly acquired terminal interface we could view and edit data on the screen providing it was stored independently of any program package. Our data are currently stored on disk files which we use daily to enter new data, conduct analyses, and generate reports. Magnetic tape is used as a backup system. A variety of packages such as mapping programs (GIPSY, SYMAP) and other statistical programs (BMDP, CLUSTAN) provides us with additional software capabilities. SPSS is still used as an analytical and management tool. One advantage of the SPSS package is that the computer center continuously upgrades the system with new versions as they become available.

Although disk storage space has not been difficult to obtain in the past, there is the possibility of adding 20,000 more sites to the SARG data base in the not too distant future. This would surely require continued funding and approximately five times more the amount of space than we are currently allotted.

DATA ANALYSIS

It is beyond the scope of this paper to evaluate the substantive results of SARG within the past

decade. Anyone interested in these topics may consult the two main SARG publications (Gumerman 1971; Euler and Gumerman 1978). What we will discuss, however, are the analytical procedures in terms of computer applications.

Between 1971 and 1976, SARG members requested information on subsets of variables, whole data sets, or pooled data sets. The results of these individual efforts and a 5 day workshop held in 1976 culminated in the 1978 publication. Criticisms and evaluations of the early SARG project were taken into consideration in the current phase of the research. Issues included such topics as data comparability, standardization, and scale (Sullivan and Schiffer 1978).

Currently, analysis is performed in one of two ways. The first is initiated by the individual SARG member who requests a specific type of analysis such as rank-size, nearest neighbor analysis, chi-square tests, analysis of variance, etc. These analyses are performed by the SARG staff and the results are then sent to the SARG participant. This type of analysis usually relates to the participants own study area.

The second direction is being taken by the SARG staff. It is often more expedient for the SARG staff (comprised of four individuals) to perform certain types of analyses and later confer with the entire SARG membership. By characterizing the regional environment of each project area in a standardized fashion, we hope to more precisely address the question of whether prehistoric populations in similar environments utilized similar adaptive processes. These analyses are implemented though a series of computer mapping programs and accompanying statistical tests. We are also beginning to examine new sources of data such as population curves and decadic rainfall figures, and variables in the SARG format which have not yet been exploited to their full potential.

We have high-lighted some of the problems resulting from a shared data environment. Coordinating the efforts of the participants and the SARG staff at times seems formidable. Yet the advantages clearly outweigh the problems. SARG has come a long way since its inception. Enhancements in data format, acquisition, entry, and analytical techniques have provided SARG with the capability of substantively addressing issues of cultural and environmental changes in the prehistoric plateau Southwest. None of these issues could have been addressed with just one data set.

Since it is judicious to consider what long term extensions might entail, three future enhancements may be identified. First, direct data communication of the data base by each participant from their own facility would be a considerable benefit to analysis. Participants would have to acquire appropriate terminals and communication interfaces. To protect the data base, access would be in inquiry mode only. The second enhancement involves data entry by each participant. In order to assure quality control, data entered from a remote location would be collected in a special file - the validation file - and this information would be verified by the SARG staff before it is merged with the protected information in the SARG data base. The third extension would be utilizing any of a number of new technologies. An example of this would be graphics terminals. Although currently expensive, lower costs of graphics terminals in the future may make these interfaces a potential candidate for SARG research. Graphics terminals may be used to display shapes and maps used in SARG analyses as well as in data recording. This latter use, the graphic representation of basic forms used to define some of the SARG variables, would greatly enhance comparability in data recording (Gaines and Gaines 1981).

Clearly we have many problems to solve before automated data sharing becomes a common place

approach. Issues of standardization, procedural and organizational controls, quality control of data, safeguards and security measures must be addressed. Although some archaeologists would view automated data sharing skeptically with a "wait and see attitude", most would agree that the advantages in terms of costs, time and elimination of redundancy, and the potential of an expanded data base, makes data sharing a potentially powerful research approach. Hopefully, the SARG example will continue to serve as a model for other applications and we anticipate that this type of joint research will be a future trend in archaeology.

REFERENCES CITED

Euler, Robert C. and George J. Gumerman
1978 Investigations of the Southwestern Anthropological Research Group: an experiment in archaeological cooperation. Museum of Northern Arizona, Flagstaff.

Gaines, Sylvia W.
1978 Computer applications of SARG data. In Investigations of the Southwestern Anthropological Research Group: an experiment in archaeological cooperation, edited by R.C. Euler and G.J. Gumerman, pp. 119-138. Museum of Northern Arizona, Flagstaff.

Gaines, Sylvia W. and Warren M. Gaines
1981 What future data banks have to offer archaeology. In The proceedings of the X Congreso: Union Internacional de Ciencias Prehistoricas y Protohistorias. Mexico City.

Gaines, Sylvia W. and Fred Plog
1980 Continuing research of the Southwestern Anthropological Research Group. Proposal (Grant #BNS80-04571) on file with the National Science Foundation. Washington.

Gumerman, George J.
1971 The distribution of prehistoric population aggregates. Anthropological Reports 1, Prescott College Press, Prescott.

Plog, Fred and James N. Hill
1971 Explaining variability in the distribution of sites. In The distribution of prehistoric population aggregates, edited by G.J. Gumerman, pp. 7-36. Anthropological Reports 1, Prescott College Press, Prescott.

Sullivan, Alan P. and Michael B. Schiffer
1978 A critical examination of SARG. In Investigations of the Southwestern Anthropological Research Group: an experiment in archaeological cooperation, edited by R.C. Euler and G.J. Gumerman, pp. 168-179. Museum of Northern Arizona, Flagstaff.